LINKED OPEN DATA IN THE EARTH OBSERVATION DOMAIN: THE VISION OF PROJECT LEO

M. Koubarakis

National and Kapodistrian University of Athens
Greece

ABSTRACT

Lots of Earth Observation data has become available at no charge in Europe and the US recently and there is a strong push for more open EO data. Open EO data that are currently made available by space agencies are not following the linked data paradigm. Therefore, from the perspective of a user, the EO data and other kinds of geospatial data necessary to satisfy his or her information need can only be found in different data silos, where each silo may contain only part of the needed data. Opening up these silos by publishing their contents as RDF and interlinking them with semantic connections will allow the development of data analytics applications with great environmental and financial value. This is the goal of the new European project LEO which we introduce in this paper.

1. INTRODUCTION

Lots of Earth Observation (EO) data has become available at no charge in Europe and the US recently and there is a strong push for *more open EO data*. For example, a recent paper on Landsat data use and charges by the US National Geospatial Advisory Committee - Landsat Advisory Group starts with the following overarching recommendation: "Landsat data must continue to be distributed at no cost". Similarly, the five ESA Sentinel satellites that would soon go into orbit have already adopted a fully open and free data access policy.

Linked data is a new data paradigm which studies how one can make RDF data available on the Web, and interconnect it with other data with the aim of increasing its value [1]. In the last few years, linked *geospatial* data has received attention as researchers and practitioners have started tapping the wealth of geospatial information available on the Web [2]. As a result, the *linked open data (LOD) cloud* has been rapidly populated with geospatial data some of it describing EO products (e.g., CORINE Land Cover and Urban Atlas published by project TELEIOS). The abundance of this data can prove useful to the new missions (e.g., Sentinels) as a means to increase the usability of the millions of images and EO products that are expected to be produced by these missions.

However, open EO data that are currently made available by space agencies such as ESA and NASA are *not* following the linked data paradigm. Therefore, from the perspective of a user, the EO data and other kinds of geospatial data necessary to satisfy his or her information need can only be found in different data silos, where each silo may contain only part of the needed data. *Opening up these silos* by publishing their contents as RDF and interlinking them with semantic connections will allow the development of data analytics applications with great environmental and financial value.

The European project TELEIOS[1] is the first project internationally that has introduced the linked data paradigm to the EO domain, and developed prototype applications that are based on transforming EO products into RDF, and combining them with linked geospatial data. Examples of such applications include wildfire monitoring and burnt scar mapping, semantic catalogues for EO archives, and rapid mapping. The wildfire monitoring application is available on the Web[2] and has been used operationally by government agencies in Greece in the summer fires of 2012. Recently, it has also been awarded 3rd place in the Semantic Web Challenge.

TELEIOS concentrated on developing data models, query languages, scalable query evaluation techniques, and efficient data management systems that can be used to prototype applications of linked EO data. However, developing a methodology and related software tools that support the whole lifecycle of linked open EO data (e.g., publishing, interlinking etc.) has *not* been tackled by TELEIOS. The main objective of the new European project "Linked Open Earth Observation Data for Precision Farming" (LEO) presented in this paper is to go beyond TELEIOS by designing and implementing software supporting *the whole life cycle of linked open EO data* and its combination with linked geospatial data, and by developing a precision farming application that heavily utilizes such data.

LEO brings together the two core academic partners of

Additional Authors: P. Smeros, C. Nikolaou, G. Garbis, K. Bereta, S. Gianakopoulou, K. Dogani, M. Karpathiotaki, I. Vlachopoulos, D. Savva, G. Stamoulis (National and Kapodistrian University of Athens, Greece); K. Kyzirakos, S. Manegold (Centrum Wiskunde & Informatica, Netherlands); B. Valentin (Space Applications Services, Belgium); H. Bach, F. Niggemann, P. Klug (VISTA Geowissenschaftliche Fernerkundung, Germany); W. Angermair, S. Burgstaller (PC-Agrar Informations und Beratungsdienst, Germany)

[1] http://www.earthobservatory.eu/

[2] http://papos.space.noa.gr/fend_static/

TELEIOS (National and Kapodistrian University of Athens and Stichting Centrum voor Wiskunde en Informatica), two SMEs with lots of experience with EO data and their applications (Space Application Services and VISTA) and and one industrial partner with strong Farm Management Information Systems experience (PC-Agrar). LEO has started on October 1st, 2013 and will last for two years.

The rest of the paper is organized as follows. Section 2 presents the scientific and technical objectives of LEO. Section 3 presents in more detail some of the research to be carried out in the project. Finally, Section 4 concludes the paper.

2. SCIENTIFIC AND TECHNICAL OBJECTIVES OF LEO

The detailed scientific and technical objectives of LEO are the following:

1. To capture, as precisely as possible, the life cycle of linked open EO data.

2. To develop publishing tools that transform open EO data and metadata, made available by space agencies such as ESA and NASA, from their standard formats into RDF and make it available on the LOD cloud.

3. To develop publishing tools that transform open geospatial data and metadata from their standard formats into RDF and make it available on the LOD cloud. Open geospatial data (e.g., digital maps, administrative data, environmental data, etc.) are typically used together with EO data in applications such as precision farming and are made available by public agencies as well (e.g., the Bavarian Topographical Survey for our precision farming application).

4. To develop tools that interlink open EO data sources and geospatial data sources published as RDF.

5. To develop tools for cross-platform searching, browsing and visualization of linked EO data and linked geospatial data.

6. To demonstrate the value of the developed tools by:

   (a) Performing large-scale publication and linking of open EO data from the GMES Space Component Data Access warehouse managed by ESA, and relevant geospatial datasets made available by other public bodies in Europe.

   (b) Developing a precision farming application that shows how geo-information services based on linked open EO data, linked geospatial data and specialized algorithms can contribute to an environmentally friendly increase in the efficiency of agricultural production.

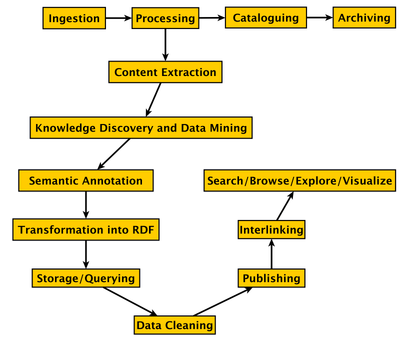

Fig. 1: The life cycle of linked open EO data

The next section discusses the first of these objectives in more detail. Because this objective refers to the whole life cycle of linked open EO data, our discussion covers most of the other objectives as well; therefore, it serves as a short introduction to the whole research agenda of LEO.

3. THE LIFE CYCLE OF LINKED OPEN EO DATA

Developing a methodology and related software tools that support the whole life cycle of linked open EO data has not been tackled by any research project in the past, although there is plenty of such work for linked data e.g., by project LOD2 and others [3, 4]. Capturing the life cycle of open EO data and the associated entities, roles and processes of public bodies making available this data is the first step in achieving LEO's main objective of bringing the linked data paradigm to EO data centers, and re-engineering the life cycle of open EO data based on this paradigm.

The life of EO data starts with its generation in the ground segment of a satellite mission. The management of this so-called payload data is an important activity of the ground segments of satellite missions. Figure 1 gives a high-level view of the life cycle of linked EO data as we envision it at this moment in project LEO (this is a preliminary version which will be further refined in the course of the project).

Let us now briefly discuss each one of these phases:

- *Ingestion, processing, cataloguing and archiving.* Raw data, often from multiple satellite missions, is ingested, processed, cataloged and archived. Processing results in the creation of various standard products (Level 1, 2, etc., in EO jargon; raw data is Level 0) together with extensive metadata describing them. For example, in the fire monitoring application developed in project TELEIOS, images from the SEVIRI sensor are processed (cropped, georeferenced and run through a pixel classification algorithm) to detect pixels that are hotspots. Then these pixels are stored as standard products in the form of shapefiles. Raw data and derived products are complemented by auxiliary data, e.g., various kinds of geospatial data such as maps, land use/land cover data, etc.

Raw data, derived products, metadata and auxiliary data are stored in various storage systems and are made available using a variety of policies depending on their volume and expected future use. For example, in the TerraSAR-X archive of DLR, long term archiving is done using a hierarchy of storage systems (including a robotic tape library), which offers batch to near-line access, while product metadata are available on-line by utilizing a relational DBMS and an object-based query language.

TELEIOS has developed two technologies that are important for the first two of the phases (ingestion and processing): the SciQL data model and query language [5] and data vaults [6]. SciQL is an SQL-based query language for scientific applications with arrays as first class citizens [5]. It allows stating complex satellite image processing functions as declarative SciQL queries, thus it eases substantially the development of processing chains run by EO data centers today. The data vault is a mechanism that provides a true symbiosis between a DBMS and existing (remote) file-based repositories such as the ones used in EO applications [6]. The data vault keeps the data in its original format and place, while at the same time enables transparent data and metadata access and analysis using the SciQL query language. SciQL and the data vault mechanism are implemented in the well-known column store MonetDB[3].

- *Content extraction, knowledge discovery and data mining, and semantic annotation.* In the DLR knowledge discovery and data mining framework developed in TELEIOS [7], traditional raw data processing has been augmented with *content extraction* methods that deal with the specificities of satellite images and derive image descriptors (e.g., texture features, spectral characteristics of the image). Knowledge discovery techniques combine image descriptors, image metadata and auxiliary data (e.g., GIS data) to determine concepts from a domain ontology (e.g., forest, lake, fire, burned area) that characterize the content of an image. Hierarchies of domain concepts are formalized using OWL *ontologies* and are used to annotate standard products. Annotations are expressed in RDF and are made available as linked data so that they can be easily combined with other publicly available linked data sources (e.g., GeoNames, OpenStreetMap, DBpedia) to allow for the expression of rich user queries.

  In TELEIOS we have experimented with implementing content extraction and KDD algorithms using SciQL instead of specialized algorithms coded in an appropriate programming language (e.g., C++ or Java).

- *Tranformation into RDF.* This phase transforms vector or raster EO data from their standard formats (e.g., shapefiles or GeoTIFF) into RDF. In LEO we will advance the state of the art in transforming EO data and geospatial data into RDF by first developing a *generic stand-alone tool* that will be able to deal with *vector data and their metadata*, and to support natively all popular geospatial data formats (shape files, KML and GeoJSON initially). The tool will produce RDF data modelled as in the recent works on stSPARQL and GeoSPARQL where new data types to encode the geometry of features have been defined. Since the transformation of raster data (e.g., raw satellite images) into RDF does not appear to be reasonable, this stand-alone tool will allow the transformation of the accompanying metadata only in such cases. As an advanced alternative, we will also *integrate the extraction and transformation functionality of the stand-alone tool into MonetDB*, a DBMS that supports both RDF (via relational mapping using the Strabon front-end developed in TELEIOS) and arrays (natively via SciQL). This approach allows the use of SciQL during the mapping process, e.g., to extract features from the raw raster data that can then be transformed into and published as RDF. Also, it opens up possibilities for on-demand extraction and transformation when querying the RDF data using the data vault machinery of MonetDB.

- *Storage/Querying.* This phase deals with storing all relevant EO data and metadata on persistent storage so they can be readily available for querying in subsequent phases. In TELEIOS, MonetDB (with SciQL and the data vault) is used for the storage of raw image data and metadata [6] while the spatiotemporal RDF store system Strabon[4] and the query language stSPARQL is used for storing/querying semantic annotations and other kinds of linked geospatial data originating from transforming EO products into RDF [8].

- *Data cleaning.* Before linked EO data is ready for publication, this step is used to clean the data by e.g., removing duplicates etc. An important issue in this phase is *entity resolution* which we discuss in more detail in the "linking" phase below.

- *Publication.* This phase makes linked EO data publicly available in the LOD cloud using well-known data repository technologies such as CKAN. In this way, others can discover and share this data and duplication of effort is avoided.

[3] http://www.monetdb.org/

[4] http://strabon.di.uoa.gr

- *Interlinking.* This is a very important phase in the linked EO data life cycle since a lot of the value of linked comes through connecting seemingly disparate data sources to each other. Up to now, there has not been much research or tools for interlinking linked EO data. If one considers other published linked data sets that are not from the EO domain, but have similar temporal and geospatial characteristics, the situation is the same. These data sets are typically linked only with `owl:sameAs` links and only to core datasets such as DBpedia or Geonames. In addition, these links are often created manually since existing tools such as Google Refine and Silk have not been found to perform satisfactorily for these datasets [9].

  In LEO we will advance the state of the art in the area of interlinking of linked open data by concentrating on the geospatial, temporal and measurement characteristics of EO data. The first problem to be studied will be entity resolution. For geospatial data, entity resolution has been studied only for location (point) datasets in [10]. We will extend the relevant techniques to the case of more complex geometries captured by the spatial literal data types of stSPARQL and GeoSPARQL that will be utilized in the data published by LEO as discussed above. If needed, we might also use ontology alignment techniques to deal with situations where the techniques of [8] would fail (e.g., when the types of features considered are synonyms or one type is a subclass of the other etc.). Finally, we will consider geospatial entity resolution among EO datasets published by LEO and already existing geospatial datasets that do not follow the stSPARQL/GeoSPARQL modelling paradigm and use different vocabularies such as W3C Geo (e.g., OpenStreetMap data published by the LinkedGeoData project). This will result in the development of techniques for geospatial entity resolution in datasets that use heterogeneous geospatial vocabularies.

  We will also study the problem of discovering other kinds of semantic links that are geospatial or temporal in nature. For example, in linked EO datasets, it will often be important to discover links involving topological relationships e.g., `A geo:sfContains F` where `A` is the area covered by a remotely sensed multispectral image `I`, `F` is a geographical feature of interest (field, lake, city etc.) and `geo:sfContains` is a topological relationship from the topology vocabulary extension of GeoSPARQL. The existence of this link might indicate that `I` is an appropriate image for studying certain properties of `F`.

- *Search/Browse/Explore/Visualize.* This is also a very important phase since it enables users to find and explore the data they need, and start developing interesting applications. For this phase in LEO, we plan to extend the tools developed in ESA project RARE[5] and the tool Sextant [11] developed in TELEIOS with additional functionalities. Finally, we plan to make these tools available for mobile devices (tablets, smartphones) to enable the use of EO data by ordinary users and application specialists alike.

4. CONCLUSIONS

We discussed the vision of the new European project LEO which intends to bring the linked open data paradigm to Earth Observation by extending the results of project TELEIOS.

5. ACKNOWLEDGMENTS

This work has been funded by the FP7 project LEO (611141).

6. REFERENCES

[1] Christian Bizer, Tom Heath, and Tim Berners-Lee, “Linked data - the story so far,” *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009.

[2] Manolis Koubarakis, Manos Karpathiotakis, Kostis Kyzirakos, Charalampos Nikolaou, and Michael Sioutis, “Data Models and Query Languages for Linked Geospatial Data,” Invited papers from 8th Reasoning Web Summer School, 2012.

[3] S. Auer, L. Bühmann, C. Dirschl, O. Erling, M. Hausenblas, R. Isele, J. Lehmann, M. Martin, P. N. Mendes, B. Van Nuffelen, C. Stadler, S. Tramp, and H. Williams, “Managing the life-cycle of linked data with the LOD2 stack,” in *ISWC*, 2012.

[4] F. Maali, R. Cyganiak, and V. Peristeras, “A publishing pipeline for linked government data,” in *ESWC*, 2012.

[5] M. L. Kersten, Y. Zhang, and M. Ivanova, “SciQL, a query language for science applications,” in *Proceedings of EDBT-ICDT, Workshop on Array Databases*, 2011.

[6] M. Ivanova, M. Kersten, and S. Manegold, “Data vaults: a symbiosis between database technology and scientific file repositories,” in *SSDBM*, 2012.

[7] D. E. Molina, S. Cui, C. O. Dumitru, M. Datcu, and Consortium members, “KDD Prototype - Phase II,” Del. 3.2.2, TELEIOS project, 2013.

[8] K. Kyzirakos, M. Karpathiotakis, and M. Koubarakis, “Strabon: A Semantic Geospatial DBMS,” in *ISWC*, 2012.

[9] P. Shvaiko, F. Farazi, V. Maltese, A. Ivanyukovich, V. Rizzi, D. Ferrari, and G. Ucelli, “Trentino government linked open geo-data: A case study,” in *ISWC*, 2012.

[10] V. Sehgal, L. Getoor, and P. Viechnicki, “Entity resolution in geospatial data integration,” in *GIS*, 2006, pp. 83–90.

[11] K. Bereta, N. Charalampos, M. Karpathiotakis, K. Kyzirakos, and M. Koubarakis, “SexTant: Visualizing Time-Evolving Linked Geospatial Data,” in *ISWC*, 2013.

[5] `http://deepenandlearn.esa.int/tiki-index.php?page=RARE\%20Project`